DOCUMENT RESUME

ED 331 889 TM 016 484

AUTHOR Johnson, Phillip L.; And Others
TITLE Correlates of Examinee Item Choice Behavior in Self-Adapted Testing.
PUB DATE Apr 91
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Algebra; *College Students; *Computer Assisted Testing; Correlation; Difficulty Level; Higher Education; Item Banks; Mathematics Anxiety; *Mathematics Tests; Self Efficacy; Statistics; Test Anxiety; Test Construction; *Test Items
IDENTIFIERS *Self Adapted Testing

ABSTRACT

The strategies examinees employ when making item difficulty level choices in self-adapted computerized testing were investigated. Subjects were 148 college students (88 females and 60 males) in an introductory statistics course. The primary instrument was a self-adapted computerized algebra test used to measure student readiness for the statistics course. Each examinee was administered 20 items from a pool of 93. Students rated their self-efficacy before the test and were administered measures of mathematics anxiety and test anxiety. Inspection of each student's data file provided an indicator of selection strategy. Examinees who chose a more difficult first test item expressed greater capability and higher confidence, reported less anxiety just prior to testing, and less anxiety about mathematics in general. When selecting additional items, examinees tended toward what was termed a sluggishly flexible strategy; they chose more difficult items after passing an item or string of items, and chose less difficult items after failing a single item or string of items. The most frequent choice was to remain at the same level. Results indicate that self-adaptive testing may be a viable alternative to computerized adaptive testing. Two figures and two tables contain data from the study. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED331889

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

_____

_____

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Correlates of Examinee Item Choice Behavior in Self-Adapted Testing

Phillip L. Johnson, Linda L. Roos,
Steven L. Wise and Barbara S. Plake
University of Nebraska–Lincoln

Paper presented at the 1991 annual meeting of the
American Educational Research Association, Chicago, IL

BEST COPY AVAILABLE

TM016484

Correlates of Examinee Item Choice Behavior in Self-Adapted Testing

The advent of item response theory (IRT) has made computerized adaptive testing a viable alternative to traditional testing methods. Rocklin and O'Donnell (1987) proposed a variant of computerized adaptive testing, called self-adapted testing, whereby examinees are allowed to choose their own items from among a number of difficulty levels. They compared examinee performance on a self-adapted test with the performances of examinees taking two fixed-item tests from the same 40-item pool. One of the fixed-item tests consisted of the 20 most difficult items, while the other consisted of the 20 easiest items. Rocklin and O'Donnell found that the self-adapted test yielded a significantly higher mean score than either of the fixed-item tests.

There are many questions surrounding self-adapted testing that have yet to be explored. The purpose of this study was to look at a number of these questions about self-adapted testing, especially those concerning the behaviors that examinees exhibit when making item difficulty level choices.

This study was exploratory in nature. The strategies that examinees employ in making difficulty level choices were of primary interest. Are examinees likely to behave in an adaptive manner? That is, will an examinee choose a less difficult item level after answering an item incorrectly and choose a more difficult item level after answering an item correctly as was suggested by Rocklin (1989)? Examinee anxiety and self-perception are important issues in testing. These issues may have an influence on examinee choice of difficulty levels. In addition, what other influences are there on item difficulty level choices?

An additional question concerns the degree of match between an examinee's ability and the difficulty of the items that he/she chooses in a self-adapted test. In computerized adaptive testing, the primary goal of the computer algorithm is to administer items to examinees that match their ability levels, resulting in efficient ability estimation. In self-adapted testing, however, examinees are free to choose whatever difficulty levels they prefer. To what degree do examinees, when administered a self-adapted test, choose items that match their difficulty levels?

Method

Subjects

The subjects were 148 students enrolled in an introductory statistics course at a large midwestern university. The subjects included 88 females and 60 males. About 20% of the subjects were graduate students and about 80% were undergraduates. Participation in the study was a requirement of the course and the results were used to determine which students were in need of remediation in basic algebra skills.

Instruments

The primary instrument used in this study was a self-adapted computerized algebra test designed to measure student readiness for an introductory statistics course. The items on the test used a four-option multiple choice format and each examinee was administered 20 items. These items were chosen from a pool of 93 items which tested basic algebra skills. Wise, Plake, Johnson & Roos (1991) provide a detailed explanation of the development of the item pool. The 93 items were ranked according to item difficulty (b) parameters and divided into eight levels of roughly equal size. Each level contained 11 or 12 items. Each time an examinee chose a given

difficulty level, an item was chosen randomly, without replacement, from that level.

The test was administered on Macintosh SE/30 microcomputers using a Hypercard program. Prior to the algebra portion of the test, each examinee was given a practice item followed by two questions designed to measure the examinee's level of self efficacy. The first of the self efficacy questions concerned perceived capability and read:

> The difficulty level of the sample item was 4 which is moderate difficulty for the items on this test. Now that you have seen the sample item, how capable do you feel that you can solve the problems on this test?

The examinees provided their answers using a 7-point scale, with 1 indicating "not capable" and 7 indicating "capable." The second question concerned perceived confidence on the test and read:

> How confident do you feel that you will do well on this test?

The options for the answer ranged from "not confident" to "confident" using a 7-point scale.

After answering each item of the algebra test, the examinees were informed of the difficulty level of the item, whether they had answered correctly or incorrectly and asked to choose the difficulty level of the next item. Since no level contained more than 12 items, examinees sometimes exhausted the items from a particular level. In those cases, examinees were asked to choose an item from another level.

Maximum-likelihood estimation was using to compute an IRT ability score for each examinee. This score was compared to a cutoff value of -.20 to determine those students requiring algebra remediation. The program also computed the testing time for each item and the standard error of ability for each examinee.

In addition to the algebra test, three other instruments were used. Each of these used a paper and pencil format. The Revised Mathematics Anxiety Rating Scale (RMARS; Plake & Parker, 1982) was used to measure examinee mathematics anxiety. The Test Anxiety Inventory (TAI; Spielberger, 1980) was used to measure student anxiety toward taking tests. Three TAI scores were used in this study: the Worry subscale, the Emotionality subscale, and the Total score. The State Anxiety Scale (Spielberger, Gorsuch, & Lushene, 1970) was administered before and after the algebra test to measure situation-specific anxiety of the examinees.

Procedure

During the first class session, students supplied demographic information, completed the RMARS and the TAI, and signed up for an algebra test administration time. The students were informed that those who obtained a low score on the algebra test would be required to attend a one hour algebra remediation session to be held during the second week of class. The students who scored below the cutoff were informed during class after the completion of all testing.

The algebra test was administered in a large room containing 10 Macintosh SE/30 microcomputers. When students arrived for testing, they were seated at a computer and asked to complete the State Anxiety Scale. Next, the examinee was given a few basic instructions concerning use of the computer and he/she started the algebra test. A medium difficulty practice item was administered followed by the capability and confidence questions. The examinee was then asked to choose the level of the first item for the algebra test. Scratch paper and pencils were provided and calculators were not allowed. No time limit was imposed during testing. Upon completion of the algebra test, the State Anxiety Scale was again administered.

Results and Discussion

Two aspects of the results warrant explanation. First, due to its exploratory nature, a substantial number of correlation coefficients were computed in this study. To perform significance tests on this many correlations would be to encourage the presence of Type I errors. To avoid this problem, the results of significance tests are not reported. Instead, the reader is encouraged to consider the magnitudes of the correlation coefficients, keeping in mind that, for 148 examinees, the standard error of the correlation will be no larger than .08. Second, the difficulty choices made by the examinees in their 20-item tests were restricted by there being only 11 items in some levels and 12 in the remaining levels. That is, after the 11th item choice, some examinees were forced to choose a difficulty level different from that which they had chosen for the first 11 items. Hence, all examinees were able to make unrestricted choices only through the first 11 items.

Descriptive Statistics for the Sample

To help orient the reader to the characteristics of the sample of examinees, means and standard deviations were calculated for a number of the variables used in this study. These descriptive statistics, shown in Table 1, contain some particularly noteworthy information. First, the group of examinees passed approximately 74% of the items administered. Second, the mean estimated ability value of .18 indicates that the group's ability level was comparable to that of the calibration sample. Third, there was not much difference between the pre-test and post-test state anxiety levels. Rocklin and O'Donnell (1987) hypothesized that the use of self-adapted testing should decrease examinee anxiety levels. In this study, however, the state anxiety levels showed a small, nonsignificant increase.

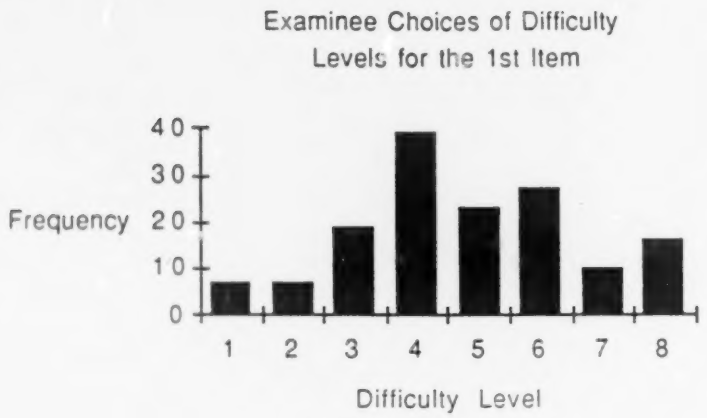
Table 1

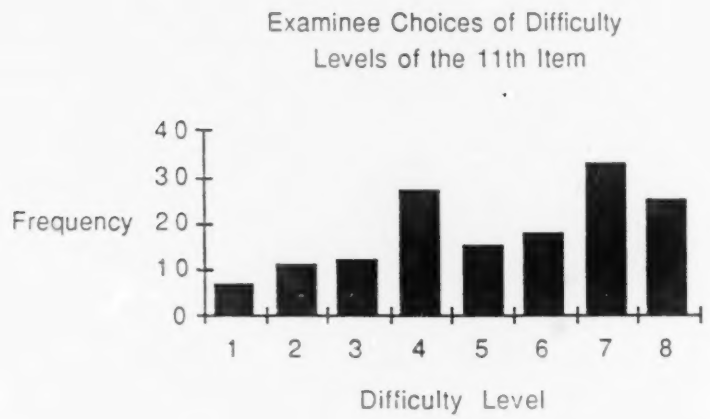
Descriptive Statistics for the Group of Examinees

Variable	Mean	Standard Deviation
Age	25.10	7.96
No. of Previous Algebra Courses	2.31	1.04
Years Since Last Algebra Course	6.64	7.62
Math Anxiety	51.70	19.41
TAI (Total)	37.24	11.57
TAI (Worry)	13.03	4.59
TAI (Emotionality)	16.13	5.10
Pre-Test State Anxiety	38.72	10.53
Post-Test State Anxiety	39.59	12.27
Number of Items Passed	14.71	2.56
Estimated Ability (Theta)	0.18	1.07
Standard Error of Ability	0.40	0.12

Distribution of Examinee Difficulty Level Choices

While all the difficulty levels were used throughout the test, the group showed a tendency toward more difficult items as the testing progressed. Figure 1 shows the distributions of difficulty level choices for the 1st, 11th, and 20th items. The mean difficulty levels for these items were 4.79, 5.32, and 5.25, respectively. Inspection of Figure 1 shows a clear shift in the distribution from the 1st item to the 11th item toward the upper categories. After the 11th item, when some examinees had exhausted a level of its items, the distribution





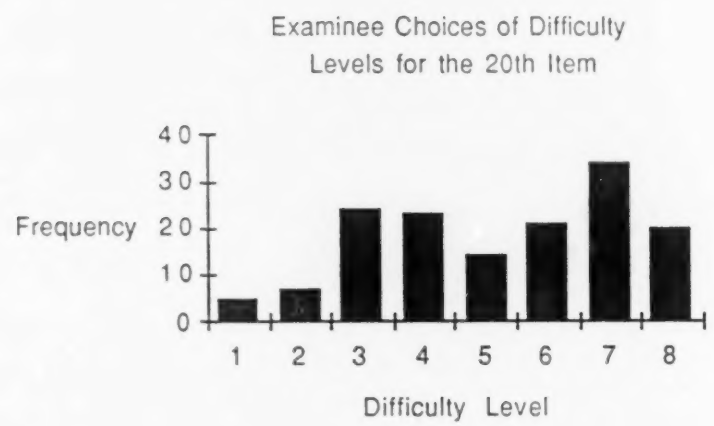


Figure 1. Distribution of difficulty levels chosen for items 1, 11, and 20.

showed a little change. The trend toward more difficult items found in this study should, however, be interpreted with caution. If the item pool had been substantially more difficult, an opposite trend may have occurred.

Correlates of the First Item Choice

Which variables were related to the examinee's choices of initial difficulty level? Table 2 shows the correlations between the first item choice and eight other variables measured in the study. The strongest correlate was perceived capability followed by perceived confidence. Math anxiety and pre-test state anxiety showed moderate correlations, and test anxiety (TAI) and number of

Table 2

Correlates of the Examinee's Choices of Their First Item Difficulty Levels

Variable	Correlation
TAI (Total)	-0.18
TAI (Worry)	-0.12
TAI (Emotionality)	-0.19
Math Anxiety	-0.48
Pre-Test State Anxiety	-0.46
Years Since Last Algebra Course	-0.16
Percieved Capability	0.85
Perceived Confidence	0.74

years since last algebra course showed relatively weak correlations. To summarize the directions of the stronger correlations, examinees choosing a more difficult first item expressed greater capability and higher confidence, while reporting less anxiety just prior to testing and to math in general.

Although the correlates of the first item choice were not surprising, the correlations of capability and confidence on subsequent choices were interesting. For capability, the correlations for the 1st, 11th, and 20th difficulty levels were .85, .69, and .53, respectively. For confidence, the correlations were .74, .63, and .46, respectively. These decreasing correlations of capability and confidence with item choice are likely due to the feedback provided by earlier test items exerting an increasing influence on difficulty level choice.

Strategies of Item Selection

In self-adapted testing, examinees may choose a variety of strategies for selecting the difficulty levels of their items. Rocklin (1989) proposed three common strategies: (1) "flexible," in which examinees choose a higher level after passing an item and choose a lower level after failing an item, (2) "failure tolerant," in which examinees choose a higher level when they pass an item and choose the same level when they fail an item, and (3) "failure intolerant," in which a lower level is chosen failing an item and the same level is chosen after passing an item. Identification of these strategies, however, is complicated by bounded nature of the difficulty level choices. That is, an examinee who passes an item at the highest difficulty level cannot use either a flexible or a failure tolerant strategy because there are no higher levels to choose.

In this study, inspection of the data file was used to attempt to describe the strategies that examinees employed. From this analysis, some tentative inferences can be drawn. First, few (if any) examinees showed a strong adherence to a flexible, failure tolerant, or a failure intolerant strategy. Instead,

most examinees exhibited what might be termed a "sluggishly flexible" strategy. In this strategy, examinees chose more difficult items after either passing an item or passing a string of several items. Likewise, these examinees moved to a less difficult level after they failed a single item or a set of several items. Instances of examinees choosing a lower difficulty level after passing an item or choosing a higher difficulty level after failing an item were rare. Figure 2 illustrates this point; it shows, broken down by whether the previous item had been passed or failed, the difficulty level choices made by examinees for items 2-11. Note that, regardless of whether the previous item was passed or failed, the most frequent choice was to remain at the same difficulty level.

		Score on Previous Item	
		Right	Wrong
Difficulty Level Chosen For Next Item	Higher	306	18
	Same	744	195
	Lower	26	191

Figure 2. Frequencies of relative difficulty level choices depending on whether the answer to the previous item was right or wrong. The frequencies are based on choices 2-11, taken across examinees.

There were some examinees who inflexibly chose either the easiest or the most difficult items available. That is, they chose all of the level 1 (or level 8) items until none were left, and then chose level 2 (or level 7) items for the remainder of the test. It is possible that these examinees would have also exhibited a sluggishly flexible strategy if there had been easier or more difficult items in the pool. Further research should be directed toward understanding the choice behavior of these examinees.

Match Between Items Chosen and Ability

Possibly the most interesting results found in this study concern the relationships between the difficulty levels chosen and examinee ability level. Note that, according to item response theory, the location of an examinee's ability estimate is not dependent upon the items administered. Hence, an examinee should expect the same ability estimate regardless of which items he/she chose. In computerized adaptive testing, the computer algorithm's goal is to administer items whose difficulty levels match an examinee's ability. In this study, examinees tended to choose items whose difficulty levels were strongly matched to their ability levels. The correlations between ability and the difficulty levels chosen for the first 11 items ranged from .64 (item 1) to .83 (item 11). The magnitudes of these correlations indicate that, in self-adaptive testing, examinees are aware of how difficult their items should be. This implies a metacognitive awareness on the part of the examinees that warrants further study.

Conclusions

The goal of this study was to investigate how examinees behave when taking a self-adaptive test, and to explore correlates of examinee test-taking behavior. The results of this study indicate that self-adaptive testing may represent a viable alternative to computerized adaptive testing. In addition, the

findings of Wise et al. (1991) suggest that examinees may perform better on a self-adaptive test than on a computerized adaptive test. The capability of self-adaptive testing to take into account an examinee's knowledge of his/her affective and motivational state may be important for a valid assessment of examinee ability.

References

Plake, B S., & Parker, C. S. (1982). The development and validation of a revised version of the mathematics anxiety rating scale. Educational and Psychological Measurement, 42, 551-557.

Rocklin, T. (1989, March). Individual differences in item selection in computerized self adapted testing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Rocklin, T., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. Journal of Educational Psychology, 79, 315-319.

Spielberger, C. D. (1980). Preliminary professional manual for the Test Anxiety Inventory. Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). Manual for the State-Trait Anxiety Inventory. Palo Alto, CA: Consulting Psychologists Press.

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1991, April). A comparison of self-adapted and computer-adaptive tests. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.